Predicting the Understandability of Imperfect English Captions for People Who Are Deaf or Hard of Hearing

SUSHANT KAFLE and MATT HUENERFAUTH, Rochester Institute of Technology

Automatic Speech Recognition (ASR) technology has seen major advancements in its accuracy and speed in recent years, making it a possible mechanism for supporting communication between people who are Deaf or Hard-of-Hearing (DHH) and their hearing peers. However, state-of-the-art ASR technology is still imperfect in many realistic settings. Researchers who evaluate ASR performance often focus on improving the Word Error Rate (WER) metric, but it has been found to have little correlation with human-subject performance for many applications. This article describes and evaluates several new captioning-focused evaluation metrics for predicting the impact of ASR errors on the understandability of automatically generated captions for people who are DHH. Through experimental studies with DHH users, we have found that our new metric (based on word-importance and semantic-difference scoring) is more closely correlated with DHH user's judgements of caption quality—as compared to pre-existing metrics for ASR evaluation.

CCS Concepts: • **Human-centered computing → Empirical studies in accessibility**; **Accessibility design and evaluation methods**;

Additional Key Words and Phrases: Accessibility for people who are deaf or hard-of-hearing, automatic speech recognition, real-time captioning system, caption understandability evaluation

ACM Reference format:
Sushant Kafle and Matt Huenerfauth. 2019. Predicting the Understandability of Imperfect English Captions for People Who Are Deaf or Hard of Hearing. *ACM Trans. Access. Comput.* 12, 2, Article 7 (June 2019), 32 pages.
https://doi.org/10.1145/3325862

1 INTRODUCTION

Many Deaf and Hard-of-Hearing (DHH) individuals across the globe benefit from offline captioning (e.g., for pre-recorded television programming) or real-time captioning services (e.g., in classrooms, meetings, and live events). With the recent improvements in the accuracy and speed of automatic speech recognition (ASR), many ASR-based applications are now seeing wide commercial use, in a variety of consumer applications. Due to their low cost and scalability, ASR systems have great potential for the task of automatic captioning [2, 16, 17]. Broadly, our research

This material was based on work supported by a Google Faculty Research Award and by the National Technical Institute for the Deaf (NTID).
Authors' addresses: S. Kafle and M. Huenerfauth, Rochester Institute of Technology, 152 Lomb Memorial Drive, Rochester, NY 14623, USA; emails: sushant@mail.rit.edu, matt.huenerfauth@rit.edu.

1936-7228/2019/06-ART7 $15.00
https://doi.org/10.1145/3325862

investigates the use of ASR technology to provide captioning services for DHH users, especially in real-time contexts such as meetings with hearing colleagues. This article investigates the design and evaluation of metrics for predicting the quality of an ASR system, to determine whether its output would produce understandable captions for DHH users.

This article is an extended version of a paper originally presented at the 2017 *ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'17)* [28], which proposed a new captioning-focused evaluation metric, called the ACE metric, for predicting the quality of automatically generated texts (from ASR) for use in captioning for DHH users. Our original conference paper provided a comparison of the ACE metric with the ubiquitous Word Error Rate (WER) metric, for a caption-understandability prediction task. Expanding upon this original paper, the current article contains additional analysis of the validity of this metric, along with comparison of its performance against other previously published metrics for evaluating the quality of ASR text output. Most substantially, this article addresses several limitations of our original ACE metric, especially in regard to how it predicts the importance of individual words and how the severity of individual ASR errors in a text can be aggregated into a composite score for an entire sentence. Finally, through a new user-based study, we have collected additional caption-preference data from DHH users, and we have performed a final evaluation of the performance of these metrics. Stimuli and study data are available in the online Appendix for this article on the ACM Digital Library and are available at http://latlab.ist.rit.edu/ace2taccess.

1.1 Automated Captioning Technologies

Real-time captioning services convert a spoken message into digital text that can be displayed, processed, represented, and stored in various forms, making it especially useful in various scenarios such as classrooms and meetings [9, 55, 64]. For people who are DHH, these services are most commonly provided by a human professional who transcribes the human-speech audio or other sounds to text using a keyboard, with the captions being displayed on a screen. Although well-trained transcriptionists can produce accurate real-time captions with a speed of over 200 words per minute (WPM) [36], systems that rely on trained transcriptionists, e.g., Computer-Aided Access Real-Time (CART) or similar services [55], are not suitable for impromptu meetings or extremely brief conversational interactions, given the overhead cost of arranging a transcriptionist. In recent work, a crowd-sourced technique to generate real-time captions [36] has been shown to produce accuracy comparable to that of a trained transcriptionist, with lower cost. However, the proposed system relies on crowd-workers with typing rates usually below that of experts (typically 77 WPM [36]) to deliver real-time captions. Further, the business model for providing such services is not yet clear.

Researchers have also examined architectures for providing real-time captioning using ASR systems, including: use of an ASR system to generate automated captions directly [2, 16, 17] or the use of non-experts to review the output of an ASR system [26, 57, 60]. Due to the ever-present concern of ASR accuracy in realistic settings, asking humans to correct errors in captions generated by ASR sounds promising, although it has been found that manually identifying and correcting errors in ASR generated captions can take more time than producing the transcript without the support of ASR [21]. With recent improvements in ASR, recent work has investigated the potential of ASR technologies (without any human intervention) for captioning live meetings [14] or for classroom lectures [31].

1.2 Word Error Rate as an ASR Metric

Accurate, large-vocabulary, continuous speech recognition is still considered an unsolved problem. Although there have been recent leaps in the performance of these systems [63], ASR performance

$$WER = \frac{S + D + I}{N}$$

Fig. 1. Formula for Word Error Rate (WER), which is based on S (number of erroneous substitutions of one word for another), D (number of deletions, i.e., erroneous omissions of words that were spoken), I (number of insertions of spurious words in the ASR output), and N (the number of words in the sentence that was actually spoken).

is generally not on par with humans, who currently provide most caption text for DHH users. ASR systems could produce errors due to noise in the input audio, the ambiguity of human speech, or unforeseen speaker characteristics (e.g., a strong accent). As researchers continue to improve ASR accuracy, they generally report the performance of their systems using a metric called Word Error Rate (WER) [62]. Given the popularity of this metric, it is reasonable that reducing WER may be a goal of many ASR research efforts (implicitly, if not overtly).

WER is calculated by comparing the "hypothesis text" (the output of the ASR system) to the "reference text" (what the human actually said in the audio recording). As shown in Figure 1, the metric considers the number of misrecognition mistakes in the hypothesis text, normalized by the word-length of the reference text. Notably, WER does not consider whether some words may be more important to the meaning of the message or whether some words might be more important than others in a text. This is a concerning limitation, because researchers have previously found that humans perceive different ASR errors as having different degrees of impact on a text; i.e., some errors might distort the meaning of the text more harshly than others [41]. Other researchers have found that the impact of errors may be dependent upon the specific application in which ASR is being used [15, 41].

1.3 Metric of ASR Quality for DHH Readers

The premise of our research is that rather than simply counting the number of errors, it would be better to consider which words are incorrect or where they occur in the sentence when evaluating ASR text output for captioning applications for DHH users. As discussed in Section 2, some researchers have previously examined the limits of the WER metric and have considered some alternatives. Our research is novel in that we are specifically interested in measuring the quality of ASR output for a captioning application for DHH individuals, and we evaluate our proposed metric in a user-study with DHH participants.

There are reasons to believe that it is important to create and evaluate metrics for measuring ASR output quality specifically targeted for DHH users. Consequently, some accessibility researchers have argued that ASR-generated errors on captions are more comprehension-demanding than human produced errors [2, 34]. Further, prior research has characterized differences in literacy rates and reading mechanisms between DHH readers and their hearing peers: Standardized testing in the U.S. has measured lower English literacy rates for deaf adults [28, 39]. Furthermore, literacy researchers have hypothesized that the basic mechanism employed by many deaf adults to understand written sentences differs from that of hearing readers: Previously, DHH users have been shown to have enhanced visual search and selection ability compared to hearing users [13]. While others have hypothesized that these readers may identify the most frequent content words and derive a complete representation of the meaning of the sentence, ignoring other words [5, 10, 56]. This reading strategy is often referred to as a "keyword" strategy, and it suggests that a subset of the words in a caption text might be of very high importance to DHH users (for text understandability). Following this same reasoning, it might be disadvantageous to penalize each error in a caption text equally. Some errors may be very consequential to the understandability of the text (with the potential to mislead or confuse the readers), while other errors may have

little impact (perhaps easily ignored by readers). Our goal is to develop a metric that can predict the quality of an ASR text output based on the understandability of the text as a caption for DHH users. Unlike WER, we want our metric to distinguish between harmful errors in the caption (likely to degrade the quality of caption for DHH users) and less harmful errors; the metric should use this distinction when penalizing a text for each type of error.

With this aim, we seek to identify an ASR evaluation metric that is more captioning-focused, to measure the impact of errors on the understandability of a caption for DHH users. Specifically, we investigate a new ASR evaluation metric that considers the importance of the spoken word for understanding the meaning of the spoken message—and the semantic deviation in the meaning due to each error. In the coming sections, we will discuss how we design such a metric and how we evaluate it in a caption quality evaluation study with DHH users.

2 RELATED WORK

This section surveys prior research on the limitations of WER as an evaluation metric for ASR research and prior efforts to design alternative metrics.

2.1 Limitations of WER

While WER is the most commonly used metric for evaluating speech recognition performance, researchers have argued for alternative evaluation measures that would better predict human performance on tasks that depend on ASR text output usability [40, 42]. There have also been concerns about the nature of the metric: Researchers have criticized that while WER has a lower bound of zero (indicating that a hypothesis text is a perfect match for a reference text), WER lacks a proper upper bound, making it difficult to evaluate WER scores in an absolute manner [40]. Further, researchers have also argued that WER is ideally suited to evaluation of ASR quality only for those applications in which the human can correct errors by typing, since the WER metric is based upon counting errors, which directly relates to the cost of restoring the output word sequence to the original input sequence [40].

In other applications, researchers have observed a weak relationship between WER and human task performance. For example, in the task of spoken document retrieval (in which a human is searching for a speech audio file, which has been transcribed by ASR, by typing search terms for desired information), researchers have found that the WER of the ASR system has little correlation with the retrieval system performance [20, 23]. Moreover, Wang et al. [61], saw improvements in a spoken language understanding task, even during a significant increase in WER.

2.2 Methods of ASR Evaluation

Several researchers have proposed alternative metrics to WER for evaluating the performance of ASR for specific applications. Nanjo and Kawahara [43] have weighted errors based on the Term Frequency-Inverse Document Frequency (TF-IDF) measure, in the context of a keyword-based open-domain speech understanding application. TF-IDF is commonly used by researchers studying information retrieval; it assigns high scores to words that are generally "rare" but which appear in great frequency in a particular document, e.g., if a rare word like "daffodil" appears very frequently on a particular webpage, then it is reasonable to think that the word "daffodil" is an important "keyword" for that webpage. Specifically, the researchers have used TF-IDF as a "loss function" during the decoding step of their ASR system [43]. (During decoding, the ASR system aims to determine the most likely sequence of words that corresponds to speech information.) The loss function penalized errors on keywords more heavily than errors on other words, when choosing from a list of output candidates. The authors explored using this metric as a weighting

factor in a Boolean fashion (i.e., is something a keyword or a non-keyword) or by using the actual numerical TF-IDF scores as weights.

Garofolo et al. attempted to modify WER to weight "content words" more heavily than other words [20]. Generally speaking, content words include nouns, verbs, adjectives, and adverbs that convey semantic meaning, rather than "function words," e.g., determiners, which convey grammatical information. The authors used ASR in an information retrieval application; users searched for excerpts in large spoken audio recordings. The authors found a nearly linear relationship between their proposed metric and retrieval performance across different systems; i.e., ASR systems that recognized content words more accurately provided the best input for their retrieval task. To summarize, References [20] and [43] found that keyword identification (to differentially weight specific kinds of errors) led to useful ASR metrics for applications related to information search.

Some researchers have considered applications of ASR that are even closer to our focus on automatic captioning: For instance, some have proposed a metric for evaluating ASR output on a speech transcription task [41]; their metric was based on opinion scores collected from humans who judged the quality of ASR-generated voicemail-to-text transcripts. Scores from their metric correlated with the human judgments better than WER did. Their metric learned the cost of different error types (namely, insertion, deletion, and substitution) and learned a weight factor called the "saliency index" for words to predict their contribution in text understandability. While not focused on creating a fully automatic metric, Apone et al. [1] investigated different categories of captioning errors (e.g., substitution of a word with an incorrect tense) and weighted each category to design a "weighted WER" metric. This metric was proposed for evaluating the accuracy of captions for television.

The "match error rate" (MER) and "word information loss" (WIL) metrics were introduced in Reference [42], as replacements for WER in settings where high error rates are common. The MER metric is similar to WER except that it is properly normalized and thus computes the "probability" of a given match (between the reference text and the hypothesis text) being incorrect. Similar to MER, WIL is a probabilistic approach that approximates the proportion of the word information lost due to the presence of errors.

Researchers in machine translation (MT) have proposed numerous evaluation metrics that compare the similarity of an automatically translated sentence with one or more reference sentences. At first blush, speech recognition could be considered as a translation task where speech signals are "translated" to text. From this perspective, we could consider whether it is suitable to compare the reference text and the hypothesized output from ASR by treating the reference text as the ideal translation and hypothesized output as the machine-generated translation. However, this analogy has limits. For instance, the speech recognition task is substantially different than the machine translation task: A perfect speech-recognition output must always be similar to the reference text. In contrast, a perfect machine-translation output could differ from the reference translation, while still preserving the meaning. To account for this possibility of multiple possible translations for a text, popular MT evaluation metrics (e.g., METEOR [3], BLEU [45]) consider for the non-linear relational-mappings between words in the reference text and in the hypothesized text during evaluation. Such approaches are less appropriate for ASR evaluation, which depends upon a match between the reference and hypothesis text.

Our work is also inspired by the work of McCowan et al. [40], which discussed the challenges of application-oriented evaluation of ASR systems and proposed a generic framework to evaluate the ASR output based on information retrieval concepts like "precision" and "recall." Their framework treated the speech recognition task as analogous to an information retrieval task, i.e., the goal for transcription is to retrieve all the relevant information (i.e., the spoken word) from in the original speech signal. In their framework, they provided room to incorporate application-dependent

importance weights for words and for different ASR error types. However, for our application of real-time captioning for DHH users, the assumptions made in their framework are less appropriate: They treated words as independent units of information, without considering their position in a sentence, i.e., under this assumption, identical words located at different positions in a sentence will have identical weights. Later, in Section 7, we provide a comparison of performance these metrics on a caption understandability prediction task for DHH users.

In contrast, we propose a new captioning-focused evaluation framework called the Automatic-Caption Evaluation (ACE) framework to accurately model the impact of an error in the understandability of a caption-text. To measure the impact of an error in a caption-text, the framework considers the importance of words and the semantic deviation due to the error. With the help of this framework, the article discusses the design of several caption quality evaluation metrics and provides evaluations of their performance, through studies with DHH users.

3 AUTOMATIC-CAPTION EVALUATION FRAMEWORK

As discussed in Section 1, we are interested in the potential for ASR systems to be used as a real-time captioning tool for impromptu meetings. There are many commercial and research ASR systems available, each with different capabilities, e.g., adapting to the voice of specific speakers, operating in contexts with different types of background noise, or recognizing different vocabulary or genres [25, 37, 38]. A natural question is how to compare ASR systems to determine their suitability for use in this context. Given the limitations of WER discussed in Section 2, we therefore present a new framework, called the Automatic-Caption Evaluation (ACE) framework, which helps the design of a better evaluation metrics for carefully assessing the efficacy of these tools.

The framework considers two primary factors for evaluating the impact of an error in a caption text: (a) the importance of the spoken word (reference) in understanding the meaning of the message and (b) the semantic deviation between the error word and reference word. These two factors are used to predict the impact of an error in a caption text as follows:

$$I\left(w_r,\ w_h\right) = \alpha * IMP\left(w_r\right) + (1 - \ \alpha) * D\left(w_r,\ w_h\right), \tag{1}$$

where the $(w_r,\ w_h)$ pair represents a recognition pair obtained after comparing (aligning) the automatic caption text with the actual human transcription of the spoken message, such that $(w_r \neq\ w_h)$. $IMP(w_r)$ represents the importance score of the reference word (w_r) in the meaning of the spoken message, $D(w_r,\ w_h)$ represents the semantic distance of the aligned pair $(w_r,\ w_h)$, and $I(w_r,\ w_h)$ represents the impact due to the error. Alpha (α) represents the interpolation weight, which determines how much each of the two factors (word-importance or semantic-distance) contributes to the overall impact score. In other words, the overall impact of an error is determined by the weighted combination of the importance of the reference word and the semantic distance between the error word and the reference word. The weighting factor is determined by the value of alpha (α).

It should also be noted that this framework operates on a per-error basis, meaning it considers single error at a time. However, as we will discuss in later sections, the errors caused by ASR are not always isolated to a single word, but instead, an error may be best understood as having affected an entire phrase, depending on how the alignment is performed. In the coming sections, we will provide a more detailed explanation of the key components of this framework and how we use them to create different automatic caption evaluation metrics.

3.1 Word Importance Sub-Score

The word importance measure attempts to quantify the semantic contribution of a word to the user's understanding of a text. This contribution could be based on various complex underlying

factors (e.g., parts-of-speech of the word, semantic role of the word and prior context of the conversation). This could also be a very subjective measure, which could easily be influenced by factors such as user's literacy level in the language of speech, previous experience with the topic of discussion, and so on. Furthermore, there could be challenges specific to the application of captioning impromptu meetings for DHH users (some of which are discussed briefly in Section 1.3).

Despite this complexity, several prior researchers have proposed measures of word importance in a text—for instance, statistical measures like TF-IDF are commonly used. Furthermore, inspired by the reading strategies of deaf readers (discussed in Section 1.3), we formulate other unsupervised measures of word importance, based on word predictability. Two different strategies for realizing this measure are discussed in Sections 5.1.1 and 6.1.1.

3.2 Semantic Distance Sub-Score

Misrecognition errors in automatic captioning systems may be identified by comparing the caption text with a human transcription of what was actually said by the human speaker. This comparison is typically conducted through a process called "text-alignment." The semantic distance sub-score is used to measure the quality of an aligned-unit by measuring how far a prediction is from the actual message. Notably, compared to the aforementioned word importance sub-score, the semantic distance sub-score considers the quality of the transcription itself, without regard to its importance in the context. For example, an error on an important word could be even more harmful to a text's understandability, if the erroneous word that is displayed (in place of the correct word) is especially misleading or confusing.

3.3 The Weighting Variable

The word importance sub-score (Section 3.1) and the semantic distance sub-score (Section 3.2) are combined using a weighted sum to produce an error impact score (as shown in Equation (1)), but that equation requires us to select a tuning parameter alpha (α) to specify how much each sub-score contributed to the overall error impact score.

3.3.1 Error Comprehension Dataset: Selecting the Weighting Variable. To learn the appropriate value of alpha (α), we perform a grid-search to find an optimal value for this parameter (using Equation (1)) with the help of a dataset of texts containing ASR generated errors and their estimated impact scores. In this case, we collected ASR errors and subsequent impact scores based on the response data we had collected during a prior study with DHH participants [29]; this is referred to as the "QUESTION-ANSWER" study in Figure 2. Participants were ages 20 to 32 (mean = 22.63 and deviation = 2.63), included 12 men and 18 women, and identified as DHH (26 participants self-identified as Deaf and 4 as Hard-of-Hearing). The participants were presented with imperfect English text passages (containing errors that had been produced by an ASR system when the text had been spoken aloud and then automatically recognized), and then participants were asked to answer questions that required understanding the information content of each passage. Each question was based on information from a single sentence in the passage, and each sentence contained 0 or 1 ASR errors. Each text received a score of 1 if the participant answered the question correctly and a score of 0 if the participant answered incorrectly.

From this dataset we had previously conducted, we examined the subset of question-responses that corresponded to English sentences that contained ASR errors. This data was used to calculate an aggregate "comprehension score" for each sentence, by averaging the scores from the 30 participants on questions about that sentence.

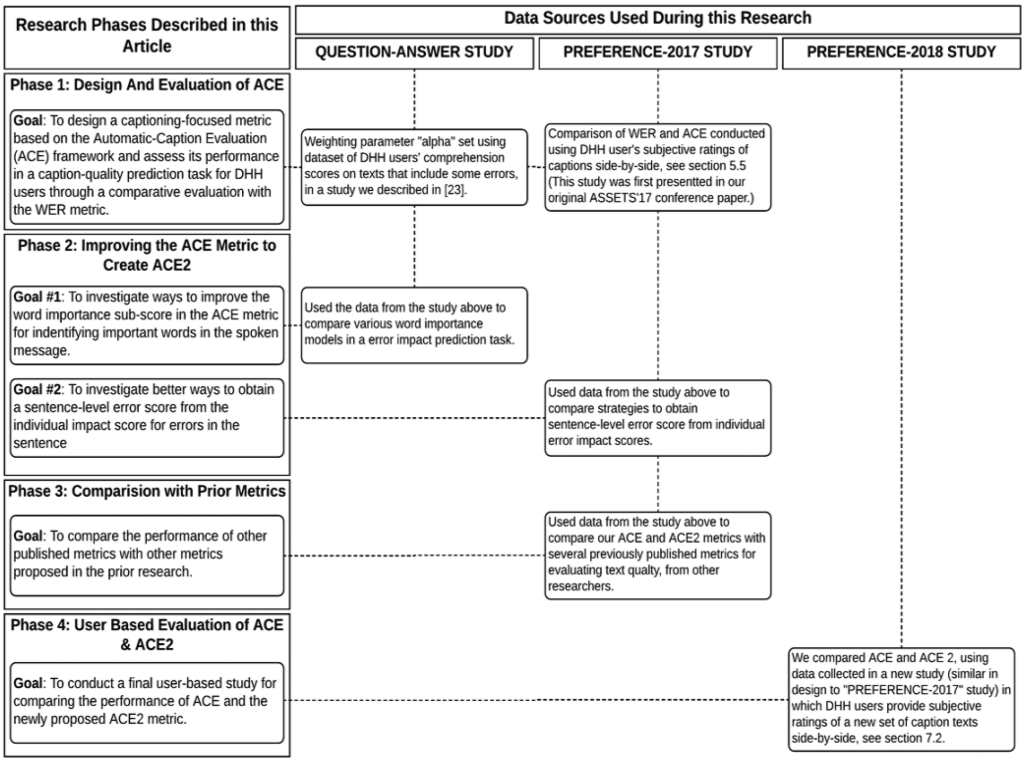


Fig. 2. Graphical illustration of research activities presented in this article.

4 RESEARCH GOALS AND METHODOLOGIES

While the goal of our overall project is to develop a metric that could automatically evaluate the understandability of a caption for DHH users, we must select a more specific scope to design a study to evaluate the efficacy of this new metric. We have selected to focus on measuring the efficacy of ASR for providing captions during a business meeting between a hearing person (speaking English) and a DHH participant. The rationale for this focus is that impromptu one-on-one meetings in a workplace may be a situation in which it is unlikely for professional captioning services to be scheduled or available. Therefore, there is an opportunity for using automatic methods like ASR technology as a captioning tool.

4.1 Four Phases of this Research

Our research methodology consists of four phases, which are described briefly below and depicted in Figure 2. Phase 1 corresponds to the work that we had previously described in our earlier conference paper [30]. Each phase is described more thoroughly in Sections 5–7.

Phase 1: We present our first metric, called the ACE metric, based on the automatic-caption evaluation framework (described in Section 3) and evaluate its efficacy in predicting the quality of automatically generated captions for DHH users. We predict DHH users will subjectively prefer ASR text output that is predicted as less erroneous by our ACE metric (as compared to ASR text output predicted as being less erroneous by the traditional WER metric). Specifically, if we ask DHH users to evaluate the quality of ASR text output, we hypothesize the following:

H1: If we compare ASR output texts predicted as better by WER (i.e., with low WER-to-ACE ratio) to ASR output predicted as better by ACE (i.e., with high WER-to-ACE ratio), then DHH participants will subjectively prefer texts preferred by ACE.

H2a: The subjective preference judgments of DHH participants on ASR output texts will correlate significantly with ACE.

H2b: There will be a significantly higher correlation between DHH human judgments and ACE, as compared to the correlation between DHH human judgments and WER.

Phase 2: We explore different ways to improve the ACE metric and evaluate different improvement strategies to propose an improved version of the metric, called the ACE2 metric.

Phase 3: We compare the performance of our metric with other published metrics in prior work on evaluating ASR text output or on caption-quality prediction for DHH users.

Phase 4: We conduct a new user-study with additional DHH users and new text stimuli, to compare our original ACE and our new ACE2 metrics, where we hypothesize the following:

H3: If we present ASR output texts with varying ACE and ACE2 scores, then there will be a significantly higher correlation between DHH human judgments and the ACE2 metric, as compared to the correlation between DHH human judgment and the original ACE metric.

5 PHASE 1: DESIGNING AND EVALUATING THE ORIGINAL ACE METRIC

In the previous sections, we discussed the ACE framework and its various components. This section describes the methodological details of the first metric we develop with the help of this framework—we call it the ACE metric. Sections 5.1–5.3 describes how different components of the metric are formulated and computed. Sections 5.4 and 5.5 provide details on how we set up a user-study for evaluating the efficacy this metric with DHH participants. Section 5.6 discusses the results of our analysis, and Section 5.7 discusses the limitations and future work.

5.1 Computing the Word Importance Sub-Score

Researchers have hypothesized that deaf readers use a sentence-understanding strategy in which they seek content words to derive a representation of sentence meaning, potentially ignoring other information, e.g., morphosyntactic relationships between words [8, 11]. Research on the eye movements of deaf readers has also revealed that deaf readers visually fixate on approximately 30% of the words in a text. The skipped words were largely determined by lexical factors, such as how frequent a word is, the length of the word, and the predictability of the word in that sentence [5]. Similarly, Rayner et al. [51] found that both the length of the word and predictability of the word in context were related to whether readers skip over a word and to the amount of time readers spent on non-skipped words. In general, highly predictable words have been shown to be read faster and skipped more often than unpredictable words by most readers [48] and especially by less-skilled readers [5].

Furthermore, word predictability has been a common theme in prior research on assessing the readability of a text or the reading comprehension skills of a participant [12, 33, 49, 50]. For instance, the "Cloze procedure" is an assessment methodology that has been around for many years, and it is one of the most common ways of evaluating both the readability of a text and the reading skills of participants. In this task, the participant is given a text with one word omitted, and they must guess the missing word. Most standardized English-language tests (e.g., TOEFL, GRE, WRAT) utilize some variation of the Cloze procedure to evaluate participants' reading skills. The predictability of a word refers to the degree to which a reader can use the context to guess the word. For example,

The _____ was barking at the mail-man.

The predictability of the word "dog" is high given the context. The context of the word is powerful enough to provide a hint as to what the word is. Conversely, consider the sentence

The meeting is scheduled on _____.

The predictability here is very low, suggesting that the readers might not be able to rely on the context to predict the word.

Given the use of word predictability in reading assessment (Cloze tests) and given the aforementioned eye-tracking research (indicating that DHH readers are more likely to skip over highly predictable words), we decided to include word predictability in our measure of word importance for evaluating the quality of captioning for DHH users.

5.1.1 Methodology: Word Predictability as Word Importance Measure. To compute the predictability score of a word, we utilized several n-gram language models; these models are based on how frequently certain sequences of words, of various length, have appeared in large collections of text. Similar models are commonly employed in word-prediction systems for text-entry applications, e.g., Reference [19]. Based on the probability score assigned to the predictions by the language model, we compute the predictability score for the word. We trained our n-gram models (n = 1 to 5) on the Switchboard [22], the English CALLHOME, and the TEDLIUM [53] corpora, which contain a total of 1.9 million unique words. These corpora were selected because they closely represent conversational speech dialogues (similar to the one-on-one meeting context in which we are considering the use of ASR for real-time captioning).

The n-gram models were used bi-directionally, to make predictions using both left and right word-sequence contexts, independently. To rank the possible word candidates using each context, a so-called "Stupid Back-off" [8] mechanism was utilized. For ranking predictions from the left context, the following scoring function was used:

$$S\left(w_i|w_{i-k+1}^{i-1}\right) = \begin{cases} \dfrac{count\left(w_{i-k+1}^{i}\right)}{count\left(w_{i-k+1}^{i-1}\right)} & \text{if } count\left(w_{i-k+1}^{i}\right) > 0 \\ \lambda\, S(w_i|w_{i-k+2}^{i-1}) & \text{otherwise,} \end{cases} \tag{2}$$

where w_m^n represents a sequence of words from position m to n and $count(w_n^m)$ fetches the count of the word sequence from the n-gram model. We utilized a value of 0.4 for lambda (λ), as recommended by Brants et al. [8]. A similar scoring function was used to rank the candidates from the right context. The predictions from both the right and left contexts were combined and then ranked for later use.

To obtain a predictability score from these predictions, we first selected the top (N_c = 20) ranked unique candidates and transformed their count probabilities to normalized probabilities (that sum to 1). For instance, in the "The meeting is scheduled on _____" example, the language model might predict various possible words, e.g., "Monday, Friday, Tuesday, and so on." The model will predict that each of these words has some probability of appearing in that context. Based on the distribution of probability among the candidates, an entropy score was calculated as follows:

$$E\,(w) = \sum_{i=0}^{N_c} -\,P\,(w_c(i)) * \mathbf{log}\,(\mathbf{P}\,(\mathbf{w_c}\,(i))), \tag{3}$$

where $E(w)$ represents the entropy of a word w (at a unique location in the text). $w_c(i)$ is a candidate of the word w predicted by the language model, $P(w_c(i))$ is the probability of the candidate $\mathrm{w_c}(i)$ as determined by the language model and N_c is the number of candidates.

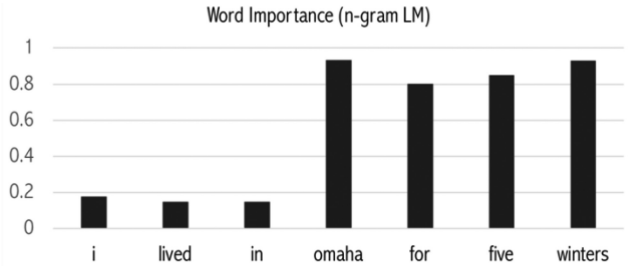


Fig. 3. Visual illustration of n-gram-based word-importance scoring, based on the predictability of words in the context of a sentence. Higher bars represent words of high importance.

The entropy score calculated in Equation (3) is a measure often used in information theory to calculate the unpredictability of a state [24]. In our application, it is the measure of the degree of unpredictability of a word given the context. A higher value indicates that the chances of picking the right word from the list of candidates are low, meaning it is difficult to predict the word. Whereas, a lower value would indicate that some words in the list of candidates clearly have a higher probability than others, meaning it is easier to predict the word. The entropy is normalized to get a predictability score within a (0, 1) range. Figure 3 provides a visual example of how this metric assigns scores to a text. As shown in the figure, some words in the example text are highly unpredictable (with higher entropy score), such as "omaha," "winters," and so on, while some are fairly predictable, such as "i," "in," and so on, in the text.

In conclusion, for each error word, the entropy score of the corresponding reference word is calculated to estimate the word predictability score. For insertion errors, where there could be two adjacent reference words that could be responsible for the error, an average entropy score is computed based on the adjacent reference words. It should be noted that our measure does not directly consider the confidence of model in predicting the "correct" word in the context, but rather looks to model the overall difficulty in making a guess using the context. This is a contrasting difference from other measures of predictability of words that have been discussed in the past, like surprisal [54].

5.2 Computing the Semantic Distance Sub-Score

The second measure that we consider is the degree to which the meaning of a word in the output text differs from the meaning of the actual word that was spoken.

5.2.1 Methodology: Vector Space Representation of Words for Semantic Distance. To compute this semantic disagreement between the error word and the actual reference word, we utilized a pre-trained word2vec[1] tool from Google. The word2vec tool provides a vector representation of a word that can be subsequently used in many natural language processing applications and research. In this article, the word2vec tool is used to compute semantic distance between two words, based on the cosine similarity of the vectors representing each word. Similar strategy of using vector representation of words for semantic distance is discussed by Frank and Willems in Reference [18].

As shown in Figure 4, each error receives a semantic distance score between 0 and 1; e.g., substitution of "winters" → "windows" receives a high semantic distance score of 0.874. For insertion and deletion errors, the length of the word is used to approximate the distance. We selected a constant value of 0.05 to scale the word length to get the semantic distance for the word. Figure 4

[1] https://code.google.com/archive/p/word2vec.

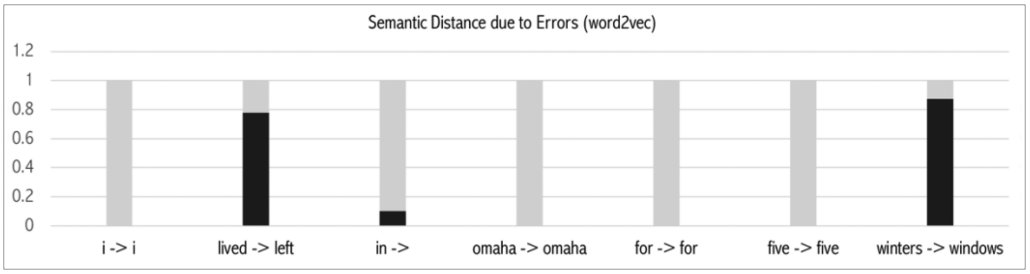


Fig. 4. Visual Illustration of word2vec-based semantic distance scoring of different alignment pairs (reference word → hypothesis word) in an example sentence. The height of the black bar indicates the semantic distance between the words.

shows that the word "in" has been deleted in the hypothesized text, thus the semantic distance due to the error is based on the length of the deleted word. This was based on an empirical evaluation on the types of ASR errors on a sample dataset: We had analyzed different ASR generated errors on 100hrs of speech from LibriSpeech [46] corpus where we found that the insertion and deletion errors usually occur on short length low-importance bearing words and rarely with longer length sentences; mean length of words inserted or deleted was 3.

5.3 From Individual-Error Impact Scores to an Overall Sentence-Error Score

So far, we have defined our methodology for computing the word-importance and semantic-distance sub-scores, which, together with the weighting variable alpha, is used to compute the impact score of an error in an automatic caption. To get the overall captioning quality score of a sentence,[2] ACE makes use of these individual impact scores due to errors in the text. As discussed in our previous work [30], there could be various ways to summarize the individual error impact scores (e.g., mean, median, etc.). However, for the ACE metric, we obtain a sentence level quality measure as follows:

$$ACE = \frac{\max\left(\mathrm{I}(w_r, w_h)\right)}{\log\left(N\right) - \mathbf{log}\left(n\right)} \tag{4}$$

where $I(w_r,\ w_h)$ represents the impact due to an error ($w_r \neq w_h$), defined in Equation (1). N is the length of the reference text, and n is the total number of errors in the hypothesized text.

With this setup, the total impact due to errors in a sentence unit is represented as the maximum score among all the individual impact scores for all errors in the sentence. The intuition here was that if a sentence contains a major error, then the overall effect on the sentence understandability can also be major. There are limitations of this approach, e.g. extending this approach to longer, multi-sentential texts; Section 5.7 enumerates some of these limitations.

As the length of the reference text increases, it slowly mitigates the impact of individual errors—the rationale being that as readers have more context (more surrounding words), it is easier to decipher the text's true meaning. But, if the number of errors increases with the reference text, then the impact of errors is counterbalanced (note the subtraction of a n term in the denominator). Using a sub-linear function (log), the rate of this change is regulated such that the effect is not always linear (e.g., the effect of large n might not be linearly reduced by a larger N). Like WER, ACE is also an error measure, meaning that a lower ACE score indicates a better caption text.

[2]A sentence in this document refers to any linguistically complete unit of text or, in the context of conversational speech text, a single unit of spoken utterance (as defined in Reference [59]).

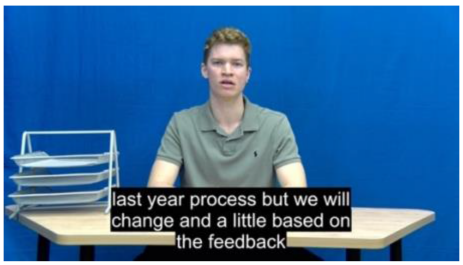


Fig. 5. Preparation of a fake meeting transcript.

Similar to WER, ACE does not have an upper bound, but it would be trivially possible to modify the metric to prevent it from exceeding some limit, *e.g.* establishing a ceiling value of 1.

In summary, our intention when designing this new ACE metric was to penalize ASR output texts that contain errors that are likely to lead to misunderstanding; specifically, the ACE metric considers errors at locations in a text that are less predictable and errors that deviate semantically from the actual word.

5.4 Designing Stimuli for Metric Evaluation (PREFERENCE-2017 Study)

We conducted a study (referred to as the "PREFERENCE-2017" study in Figure 2) with DHH participants to evaluate whether our ACE metric correlated to the subjective judgements of these users as to the overall understandability of a caption text. To create texts to display in this study, we used of some staged and prerecorded videos from colleagues at our lab [6]. These videos display one side of a two-person business meeting communication—the speaker leading the conversation in the video is made to look like he is interacting to the participant who is watching the video, as shown in Figure 5.

We extracted the verbatim script of what the human actor said during the videos, and we used the entire text as a potential source of stimuli sentences for inclusion in this study (details in Section 5.4.1). Next, we processed the original audio recording from these videos using an ASR system that we expected to make a large number of errors (it is important for our stimuli selection process described in Section 5.1.2 for us to have many possible errors to choose from). For this processing, we used the CMU Sphinx 4 system with its off-the-shelf U.S. English acoustic and language models that have been previously disseminated to the research community.

While a simplistic approach for creating stimuli for the study would have been to simply display the raw output of the ASR system to users, we were interested in obtaining judgments from participants on texts that had a variety of ACE metric scores. Furthermore, to investigate hypothesis H1, we were interested in presenting users with some pairs of ASR text output that displayed multiple hypotheses (*i.e.* two different guesses from the ASR system about what it heard), with one of the texts having a low WER-to-ACE score ratio (indicating that WER believed the text to be good, but ACE did not) and the other with a high WER-to-ACE ratio. Since ASR systems actually consider a wide variety of hypotheses when they analyze a speech audio file (with one hypothesis correct, and the remainder containing some variety of errors), we wanted to search the space of ASR output candidate hypotheses to select texts to display in our study with various WER-to-ACE ratios. Sections 5.4.1 and 5.4.2 describe our procedure for identifying ASR output hypotheses to display in our study with diverse WER-to-ACE ratios. Rather than inventing artificial errors to insert into the texts, our procedure obtains a large number of real ASR errors on a text and selects a subset of these errors to include in the texts displayed.

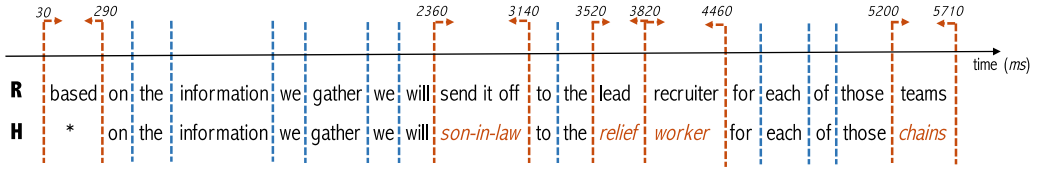


Fig. 6. Time-based alignment of reference (R) and hypothesized (H) text. The grouping with red dotted arrowhead lines indicates individualized errors aligned with corresponding reference text based on word level timestamps.

After we prepared the meeting script and ran it against our low-accuracy ASR system, the next step was to align the reference text (the verbatim script of what the human actually said) and the hypothesis text (the output of the ASR system) to obtain a list of all the errors in the ASR output. We performed a time-based alignment of the hypothesis text to the reference text to correctly identify all the errors in the hypothesis text and generated a list of confusion pairs: Each incorrectly recognized word/phrase was paired with the reference word/phrase. For additional details on our time-based alignment method, readers can refer to Reference [30].

5.4.1 Stimuli Selection. The word-alignment of the low-quality hypothesis output from the ASR system and the reference text transcripts (in Section 5.4.1 above) identified non-overlapping aligned sub-strings of these texts—with the pair of aligned substrings being non-identical if an ASR error had appeared within a particular region of the text, as shown in Figure 6.

Thus, this alignment represented a set of possible confusion pairs, with each pair corresponding to an independent error (no overlap in the time frames) the ASR system had made. We note that the reference text and the list of confusion pairs can be thought of as specifying an entire "space" of possible ASR outputs: Considering the reference text as a starting point and considering each confusion pair as an "insert an error" operator, one can imagine an entire network of possible ASR text outputs that are possible. Each ASR output contains some subset of the errors from the list of confusion pairs.

Given this space of possible ASR outputs, our goal was to identify two output texts for each reference text, with these properties:

- The output texts should reflect the reasonable performance of a commercial ASR system in noise typical of a workplace setting when the speaker is not wearing a special headset microphone; so, we wanted to identify text candidates with WER of approximately 0.25 (ranging between 20–30%); as supported by prior published results evaluating modern ASR accuracy in realistic settings [4, 35].
- We wanted to identify one text candidate that had a low WER-to-ACE ratio and another candidate with a high WER-to-ACE ratio. Specifically, we selected two candidates with identical WER: one with a high ACE score, and the other with a low ACE score.

We wrote code to execute a search procedure through the space of possibilities to identify a pair of text candidates that fit the above criteria. We executed this code on 45 sentences that had been extracted from the verbatim script of what the human spoke in our business meeting videos, and we thereby obtained 45 pairs of ASR text output candidates (two per sentence). Thus, the two text candidates identified represented two possible outputs from an ASR system. The errors that appear in the texts are realistic: They were actual errors made by an ASR system, and the overall WER error rate for the sentences is approximately 0.25. We can think of one of these text candidates as being "preferred by WER" (the one with the low WER-to-ACE ratio), and the other as being "preferred by ACE" (with the high WER-to-ACE ratio).

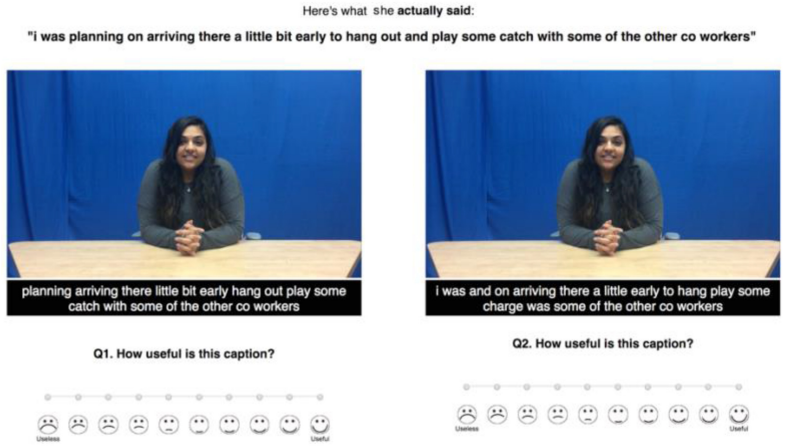


Fig. 7. Screenshot from the study, with side-by-side comparison of caption-text automatically generated by ASR. Each pair of texts (left and right) have identical WER scores, but one text in each pair was preferred by our ACE metric.

5.5 Experimental Study Setup and Procedure (PREFERENCE-2017 Study)

During the study, each participant was presented with 45 pairs of text output; each pair was displayed simultaneously, as shown in Figure 7. The reference text (what the human actually said) was provided at the top of the screen, and the two text candidates were presented on the left and right side of the screen (positioned as captioning text in a black box below the video image).

The participant was asked to provide an individual subjective quality rating for each of the two videos, using a ten-level scale (frown face to smiley face) with endpoints labeled as "Useless" and "Useful." At the beginning of the study, participants were provided with instructions on the study procedure and a practice item, prior to being presented with the 45 sentence pairs.

The WER score was identical for the two text candidates that were shown in each pair; across all 45 pairs, the WER was in the range of 0.25 to 0.3. The two versions of text differed in their ACE score; one had a higher ACE score while other had a lower score. The presentation of text candidates on the left or right side was randomized throughout the study.

We recruited participants from the Rochester Institute of Technology and surrounding campus community. We collected data from 30 DHH participants (age distribution with mean = 23.53 and standard deviation = 4.92), which included 17 men and 13 women. Among our participants, 14 people identified themselves as deaf, 8 people identified themselves as Deaf, and 8 people identified themselves as hard-of-hearing. All of the participants reported that they were familiar with the use of captioning technology, and they regularly used captioning when watching television programming.

5.6 Results and Discussion

We collected 2,700 responses in total from our 30 participants (subjective scores for each sentence, for 45 stimuli pair per participant). Figure 8 presents the average subjective judgment rating for each participant in the study, displaying their average score across all text candidates that they evaluated: the text output preferred by the ACE score (with high WER-to-ACE ratio) and the text output not preferred by ACE. Figure 9(a) presents the summarized response data across all participants, and Figure 9(b) visualizes a linear correlation best-fit line for the relationship between the ACE scores of each sentence and the participants' subjective judgment rating.

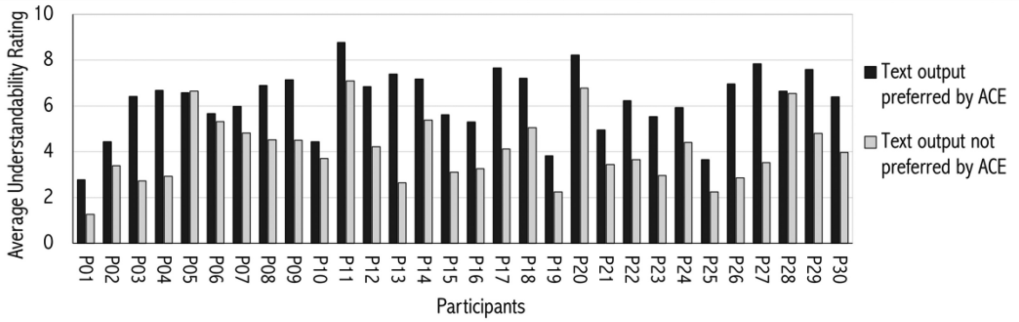
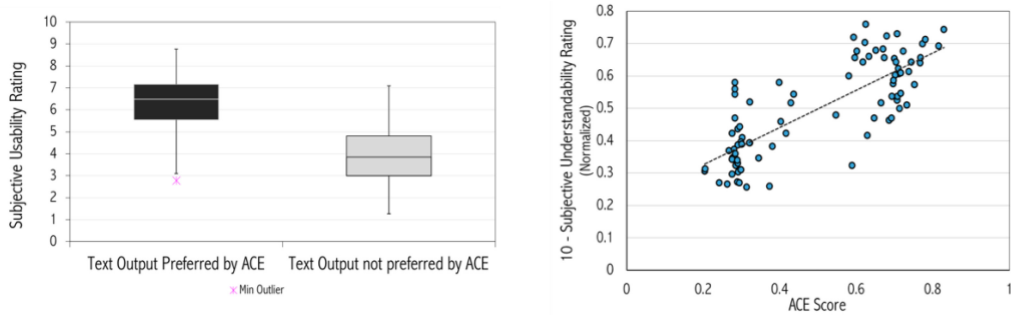


Fig. 8. Bar graph showing the average understandability rating assigned by the participants on the two versions of the text: text output preferred by ACE (black bar) and text output not preferred by ACE (grey bar).



(a) Understandability across stimuli (b) Correlation analysis

Fig. 9. Analysis of the ACE metric with participant's understandability rating.

Hypothesis H1 considered whether DHH users reported a subjective preference for captions that were predicted as "better" by our ACE metric, as compared to the captions that were predicted as "worse" by our ACE metric (high WER-to-ACE ratio vs. low WER-to-ACE ratio). The median difference (subjective score on "better" texts—subjective score on "worst" texts) was 2.5; the DHH users had higher subjective ratings for texts that had been preferred by the ACE metric. A boxplot summarizing the subjective rating scores from the participants for each stimulus (text predicted as "better" by ACE vs text predicted as "worse" by ACE) is shown in Figure 9(a). The distribution of the two groups differed significantly (Wilcoxon signed-rank W = 643394.00, N = 1350, N_test = 1226, p-Value < 0.0001). Thus, hypothesis H1 was supported: DHH users preferred predictions from the ACE metric.

For Hypothesis H2a, we considered if the understandability scores from DHH users correlated with the ACE score significantly. We computed Spearman Rho Correlation score for the two scores. The correlation coefficient was found to be rho = 0.742791 with a p-value < 0.0001. Figure 9(b) shows the corresponding correlation graph. This supports H2a: DHH user's judgments on the understandability of the caption text correlated with scores from ACE metric.

For Hypothesis H2b, we performed significant difference testing on the correlations between (1) the human subjective preferences and the ACE score (r_{ha}) and, (2) the human subjective preferences and the WER score (r_{hw}). We performed a Fisher r-to-z transformation to perform an asymptotic z-test. For r_{ha} and r_{hw}, we found a significant difference between the two coefficients (z-score = 5.771, 1-tail p-value < 0.0001). Thus, hypothesis H2b was supported: The subjective

judgment of DHH participants about the quality of ASR captions was more highly correlated with ACE, as compared to their correlation with WER.

5.7 Limitations of ACE

In our Phase 1 research, we investigated the design and evaluation of a new caption quality evaluation metric, called ACE, that analyzed the output of ASR systems to predict the impact of various ASR recognition errors on the understandability of automatically generated captions for DHH users. Further, we compared the performance of this new ACE metric to the traditional WER metric in a study with DHH participants. In a side-by-side comparison of pairs of ASR text output with identical WER score, the texts favored by our new metric were also preferred by DHH participants. Our metric also had a significantly higher correlation with DHH participants' subjective scores on caption understandability, as compared to the correlation between WER and their scores.

While we have identified word predictability and semantic distance as useful predictors of the understandability of an automatically generated caption text, there are still limitations in our metric, which we address in the next phase of this research. Some of the limitations we have identified include:

L1. The current n-gram-based word importance prediction model does not generalize well to unseen data (e.g., texts containing out-of-vocabulary words[3]), as these models are based on exact-search and match. Furthermore, other unsupervised methods of word importance scoring, such as the ubiquitous TF-IDF scoring metric, have not been fully explored.

L2. While we had identified challenges in aggregating the individual error impact scores in a text into an overall score for a sentence, we had not fully explored alternatives approaches. There could be better, more simplistic methods for calculating this sentence-level error score.

6 PHASE 2: IMPROVING THE ACE METRIC TO CREATE ACE2

The next step of our research included a closer evaluation of other unexplored strategies for building the automatic caption evaluation metric. The data collected during the PREFERENCE-2017 study in Phase 1 of our research will be utilized again in this phase of the project, to enable us to perform empirical evaluations, to explore various trade-offs between our original ACE metric and other, alternative metric designs for evaluation of ASR caption texts.

6.1 Improving the Word Importance Sub-score

To address limitation (**L1**) for the ACE metric, as discussed in Section 5.7, we explore two different approaches to improve the word importance sub-score:

6.1.1 Neural Word Predictability as Word Importance Measure. In Phase 1 of our research, we had introduced a measure of the importance of words in a spoken utterance called the word predictability score (Section 5.1), which computed the entropy of a word in its context. For the purpose, we had initially utilized several n-gram models to measure how easy (or difficult) it is to make predictions about the word given the context. However, n-gram-based models have some inherent challenges, as discussed in Section 5.7, especially when encountering out-of-vocabulary words.

To tackle these limitations, we investigated neural-network-based language models to estimate the predictability of a word in a context. Specifically, we utilize a bi-directional Recurrent Neural

[3]Words in a text that had not been present in the texts upon which the n-gram models were trained.

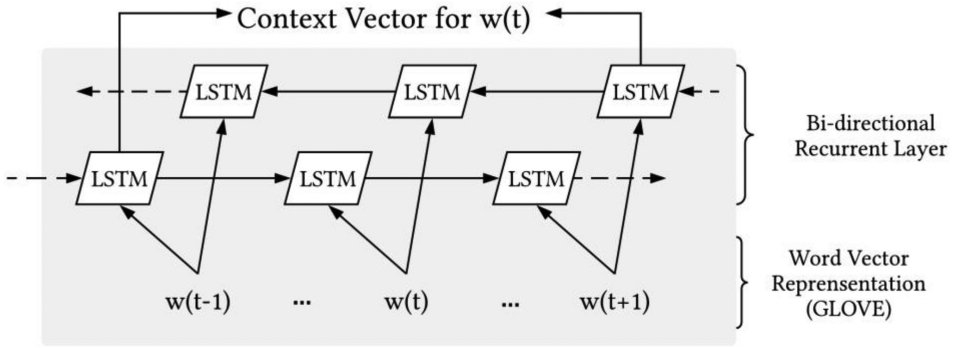


Fig. 10. Diagram of neural word predictability model demonstrating how the context of a word $w(t)$ is captured using bi-directional recurrent units.

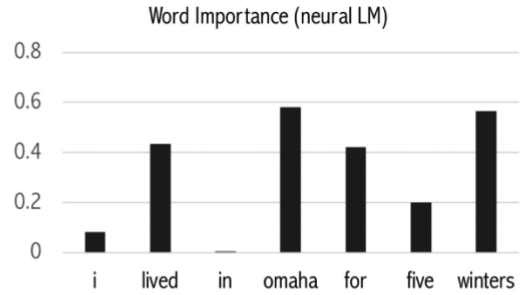


Fig. 11. Visual illustration of neural-network-based importance scoring of words based on the predictability of each word given the surrounding sentence context.

Network (RNN) architecture to build our language model: Our model uses pre-trained GLoVE[4] embedding representation for words as input, which is then processed by Long-short Term Memory (LSTM) units for context modeling.

For each word, two LSTM units are used to capture its context from both directions. To make the prediction for a word, we make use of the hidden representation from the LSTM units, as shown in Figure 10. This setup for language modeling has been discussed previously by Rei [52]. For additional details on training this neural-network-based model, refer to Appendix Section A. As of now, the model operates on utterance units (without maintaining previous dialogue contexts), and entropy of the word is then computed based on the probabilities assigned by the model on its predictions, as defined in Equation (3). Figure 11 shows how the neural predictability model assigns entropy scores of words in an example sentence.

6.1.2 Term Frequency - Inverse Document Frequency as Word Importance Measure. One of the most common techniques for identifying important words (keywords) in a text is the TF-IDF measure. This strategy is often used to identify relevant words in a document, which is based on observation from a larger collection of documents. Much like the word predictability score, TF-IDF scoring is also an unsupervised measure, thereby eliminating the need to collect subjective scores from humans, which can be both resource-intensive and time-consuming. Further, this

[4]GLoVE embedding is similar to word2vec model, discussed in Section 5.2.1, in that both approaches learn geometrical vectors for words. Our rationale for selecting GLoVE, rather than word2vec, was based on its superior performance on our dataset. The pre-trained GLoVe embedding used in our analysis was obtained from https://nlp.stanford.edu/projects/glove/.

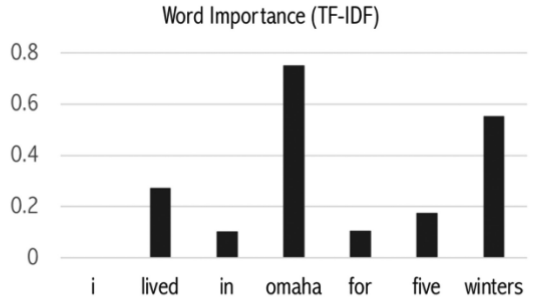


Fig. 12. Visual illustration of TF-IDF-based importance scoring of words in an example sentence.

strategy of statistically identifying importance of words in a text has been utilized in prior work on ASR evaluation [41, 43]. The TF-IDF score for a word (w), referred to as a "term," in a text document (D_w), is computed as follows:

$$tfidf(w,\ D_w) =\ tf(w,\ D_w) * idf(w,\ \boldsymbol{D}), \tag{5}$$

$$idf(w,\ \boldsymbol{D}) = \log\left(\frac{|\boldsymbol{D}|}{n}\right) \tag{6}$$

where, $tf(w,\ D_w)$ is the term-frequency, which measures the frequency of a term (w) in a document (D_w), and $idf(w,\ \boldsymbol{D})$ is the inverse-document frequency, which quantifies the presence of the term (w) in the collection document ($\boldsymbol{D}$), where $|\boldsymbol{D}|$ is the number of reference documents used and n is the number of document that contains the word w.

We utilized the Switchboard corpus [22] as our reference corpus ($\boldsymbol{D}$) for computing inverse document frequency of a word. Each two-person conversation in the corpus is treated as a "document," thus the corpus contains a total of 2,438 documents. To compute the importance of a word (w) in a dialogue, we treat the corresponding dialogue text as the observed document (D_w) and calculate the TF-IDF score of the word (w) in the document (D_w). Figure 12 shows how our TF-IDF-based word importance model scores words in an example sentence. Notably, keywords like "omaha" or "winters" are scored higher than more common words like "in," "for," and so on.

6.1.3 Evaluating the Word Importance Models. To compare these alternative methods of word-importance modeling, we compared our original ACE metric (which had used n-gram word-importance modeling) to two other 'pseudo-ACE' metrics, in which the word-importance sub-score was instead implemented using the neural-network-based or TF-IDF-based approached described above. For this analysis, we compared each metric on the data collected in our QUESTION-ANSWER study (described in Section 3.3.1), in which DHH participants answered comprehension questions for sentences containing errors.

Table 1 summarizes the performance of the candidate models on an error impact prediction task: "n-gram_LM" represents the n-gram-based word predictability measure as the word importance model (described in Section 5.1), "Neural_LM" is the neural language model-based word importance model (described in Section 6.1.1) and "TFIDF" is the TF-IDF-based word importance model (described in Section 6.1.2), and finally "word2vec" represents the word2vec-based semantic distance model (described in Section 5.2).

For the Neural_LM-based and TFIDF-based models shown in Table 1, we recalculated the optimal "alpha" coefficient for the weighted sum, as discussed in Section 3.3.1, with alpha = 0.64 for the Neural_LM-based metric and alpha = 0.48 for the TFIDF-based metric.

Table 1. Comparison of Error-impact Prediction Models (Based on Three Different Word-Importance Models) for Predicting the Comprehensibility of Error-containing Texts for DHH Users

Word Importance Model	Semantic Distance Model	MSE Loss
ngram_LM	word2vec	0.317
Neural_LM	**word2vec**	**0.255**
TFIDF	word2vec	0.257

Note: The n-gram_LM model corresponds to our original ACE metric from Phase 1.

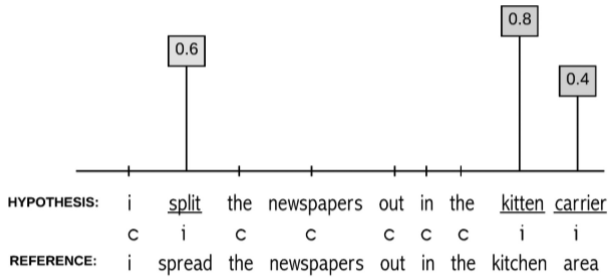


Fig. 13. An example of error impact scoring in a sentence where "c" indicates that the word was correctly recognized, and "i" indicates that the recognized word was incorrect. Recognition of word "kitchen" as "kitten" received the highest error score of 0.8 in the sentence.

For our evaluation, we considered the Mean Square Error (MSE) loss for the models, which measures the deviation of the prediction of the model with the actual error impact score received from the participants. As shown in Table 1, the neural-network-based language model was the best performing error impact prediction model, with the lowest MSE. This model out-performed our original ACE metric on this data from the QUESTION-ANSWER study. Thus, we shall use this new Neural_LM-based metric during the analysis in Section 6.2 below, and the Neural-LM word-importance model will be utilized in our new ACE2 metric, promised in Section 6.

6.2 Alternatives for Combining Individual Error Scores into a Sentence Score

Calculating an overall score for a text, based on the individual error impact scores (for specific errors that occur within the text) can be challenging, because our automatic-caption evaluation framework calculates the impact score for an error—as if it had been the only error the sentence. This approach makes it difficult to model cascading effects on comprehensibility of a text, due to multiple errors; e.g., the effect of one error may affect (add to) the effect of another. For example, in Figure 13, the impact score due to error "carrier" (for actual word "area") does not consider that the previous word "kitchen" has also been misrecognized, when making the evaluation on the impact of the error. Since "kitchen" was misrecognized, the reader loses an important context clue as to the meaning, thereby making the word "area" even more important than before.

When designing our original ACE metric in Phase 1 of this research, we had contemplated a few possible methods for calculating a sentence-level score (given the scores for individual errors within the text) as discussed in Section 5.3, but we had not systematically compared a variety of methods for this calculation (a limitation of our work that we discussed in Section 5.7, see **L2**). The sentence-level score calculation strategy used in our original ACE metric would assign the transcript in Figure 13 an error score of 0.72. In this section, we propose few other aggregation

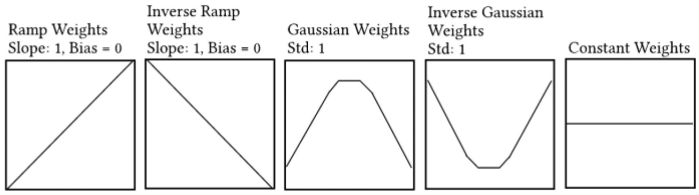


Fig. 14. Various position-based weighting functions we considered.

strategies and evaluate their performance using the data collected in the PREFERENCE-2017 study in Phase 1 of our work.

Below, we discuss some of the alternative aggregating functions that we consider for our evaluation:

(a) Mean, Median and Max: The function takes the arithmetic mean of the individual impact scores due to errors in a text. It should be noted that this function does not consider the number of correctly recognized words when making the evaluation, but only considers the number of errors (*num_errors*) and their respective impact scores $I(w_r,\ w_h)$. Thus, the *ACE_mean* metric below does not consider the number of correctly recognized words when making its evaluation:

$$ACE\_mean = \frac{\sum I\,(w_r,\ w_h)}{num\_errors} \tag{7}$$

Similarly, we consider using a median or the max function (to obtain the median among the error scores or the highest error score, respectively). For the example in Figure 13, the sentence score using these methods (mean, median, and max) would give results (0.6, 0.6, and 0.8), respectively.

(b) Position-based Weighted Average: We also consider weighting the impact scores based on the position of the words; a similar idea was discussed by Itoh et al. [27]. The intuition is that position may influence error severity, e.g., errors at the end of the text might be more prominent than those at the start. We formulated five different weighting schemes, defined by the following distributions: Ramp Weight Distribution, Inverse Ramp Weight Distribution, Gaussian Weight Distribution, Inverse Gaussian Weight Distribution, and Constant Weight Distribution.

Figure 14 illustrates the positional weighting functions. For instance, the Ramp weight distribution assigns greatest weight to the word-positions near the end of the sentence. Equation (10) shows how the weighting functions are used to combine the impact scores into a sentence score,

$$ACE_{pos_{weighted}} = \frac{\sum_{i=1}^{N} W\,(i) * I\,(w_r\,(i)\,,\ w_h\,(i))}{N}, \tag{8}$$

where N is the total number of alignment pairs during the comparison of the reference text and the hypothesis text. i represents the position of the alignment in the text and $W(i)$ represents the weight of the impact due to error based on its position i.

(c) Error-spread Model: Rather than treating an error impact score as an independent score representing the quality of a transcribed unit (or word), we wanted to consider the actual region of influence of an error in the text, when generating a quality score for the text. To realize this, we invented the Error-spread model, which essentially "spreads" the impact of an error on to its nearby words. Ideally, such a region of spread could be linguistically informed (e.g., based on semantic boundaries); however, in our initial investigation, we represent this as a constant parameter (learned) for all errors. Figure 15 shows how the error-spread model operates on the individual error impact scores from the same example sentence as in Figure 13, as we calculate an overall sentence score.

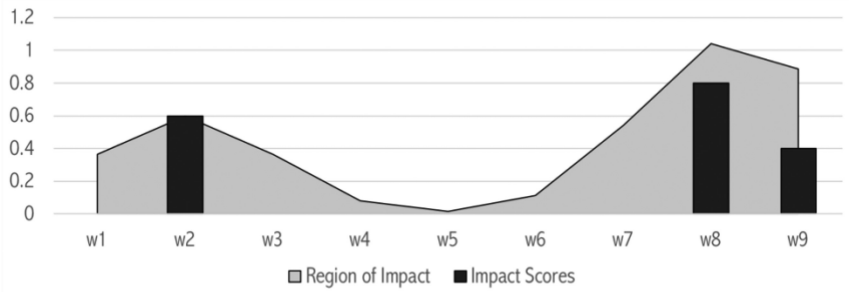


Fig. 15. This figure corresponds to the example sentence in Figure 13, and it displays a plot of impact scores for each error (black bars) and the region of impact due to error (overlaying grey region) represented using the error-spread model.

The figure illustrates how the impact of the misrecognition of the word "spread" as "split" indeed influences the nearby words, i.e., "i," "the," and "newspapers." Notably, the spread of the gray region visualizes the segment of the text primarily affected by these errors.

To implement this, for each aligned pair $(w_r(i),\ w_h(i))$representing an error $(w_r(i) \neq\ w_h(i))$, we use one-dimensional Gaussian model centered at i with a fixed standard deviation σ (defining the spread of errors to the nearby words), and the height controlled by the impact score of the error $I(w_r(i),\ w_h(i))$, to get the an error impact distribution function $I\_dist(w_r(i),\ w_h(i))$. With this representation, the region of impact is represented by the sum of all impact distribution functions. Equations (9) and (10) show how the final caption quality score is estimated, with the help of these error impact distribution functions:

$$I_{dist(w_r(i),\ w_h(i))} = I\left(w_r\left(i\right),\ w_h\left(i\right)\right) * \exp\left(\frac{-(x-i)^2}{2 *\ \sigma}\right), \tag{9}$$

$$ACE\_error\_spread = \frac{\sum_{i=1}^{N} I\_dist\left(w_r\left(i\right),\ w_h\left(i\right)\right)}{N} \tag{10}$$

Finally, now that we have enumerated various methods for aggregating individual error impact scores to calculate a sentence level score, we evaluate the performance of the various strategies. For this analysis, we have utilized the data from the PREFERENCE-2017 study in Phase 1, where DHH users subjectively rated the quality of various automatically generated caption texts. Table 2 summarizes the results of our analysis. In the table: Group (a) shows the common statistical approaches mean, median and max. Group (b) shows the position-based weighting approaches. Group (c) is the Error-spread model and, last, (d) is the aggregation function used in the original ACE metric. However, it should be noted that the error impact model used in every row of this table is the new Neural_LM-based model, which was the best in our analysis in Section 6.1.3.

Thus, we see that the aggregation approach based on the Error-spread method achieved the highest correlation with DHH users' subjective judgements of sentence quality. Based on this analysis, we can now define our new (and improved) version of the automatic-caption evaluation metric, which we shall refer to as "ACE2" for the remainder of this article. This new version of the metric consists of:

- The word-importance sub-score of the model is implemented using the Neural_LM approach described in Section 5.3.
- The semantic distance sub-score of the model is identical to the word2vec-based approach used in our original ACE metric.

Table 2. Comparison of Different Methods for Calculating a Sentence Score, Based on Individual Error Scores Contained within the Text

	Aggregate Functions	Spearman's Rho
(a)	Mean	0.797
	Median	0.719
	Max	0.837
(b)	Ramp	0.854
	Inverse Ramp	0.705
	Gaussian	0.851
	Inverse Gaussian	0.851
	Constant	0.859
(c)	Error-spread	**0.866**
(d)	Neural_LM-based error impact model using the aggregation method from the original ACE metric (see Section 5.3)	0.861

Each model below utilizes the neural_LM-based error impact model (for calculating individual error scores) discussed in Section 6.1.3.

- As discussed in Section 5.3, the new "alpha" coefficient used to calculate the weighted sum of the two sub-scores above was re-estimated, and it now has a value of 0.64.
- To calculate a sentence-level score based on the individual error impact scores, we utilize the Error-spread approach described above.

7 PHASE 3: COMPARISION WITH PRIOR METRICS

The formative evaluations presented above in Phase 2 of our research have largely consisted of comparisons among various alternative metric designs that we proposed, which has enabled us to invent the ACE2 metric defined above. In contrast, in Phase 3, we identify various metrics for evaluating ASR output text or caption-quality that have been created by other researchers in prior work. We make use of our data collected during the PREFERENCE-2017 study to compare our original ACE and new ACE2 metrics to these prior metrics. To begin, we have conducted a literature search to identify several other metrics that have been shown to be successful in predicting the quality of the ASR transcription for various applications:

—**Human Perceived Accuracy (HPA)** [41]: The HPA metric is designed to predict the human-perceived accuracy of ASR systems; it uses learned weights to differentially penalize different error types, namely insertion, deletion and substitution errors. The metric also uses a measurement called "word saliency," which measures the semantic significance (or importance) of the spoken words. Upon setting up the metric for the captioning application (by learning the errors weights using our collection of data), we evaluated its performance in the caption understandability prediction task. This metric had a correlation of 0.730 (Spearman's rho) with the DHH participants' judgements in our PREFERENCE-2017 study data, from Phase 1 of our research.

—**Information Retrieval-based Evaluation Metrics** [40]: The general idea of this metric is to treat an automatic transcription as an information retrieval task, where the goal of a transcription is to retrieve all of the spoken messages as the output. Given this portrayal, researchers suggested using standard information-retrieval-based evaluation measures like "precision," "recall," and "F-score" to report an ASR system's performance. Below, we share results using two variations of this metric—displaying Spearman rho correlations with the DHH participants' judgements in our PREFERENCE-2017 study data:

- F-score (macro-averaged): rho = 0.418
- F-score (micro-averaged): rho = 0.778

—**Word Information Lost (WIL)** [42]: This metric measures the proportion of word information communicated (or, inversely, lost). It is based on Mutual Information (MI), which measures the statistical dependence between the input words and output words. However, WIL provides a simple easy-to-implement probabilistic interpretation to the MI-based theory. The WIL metric had a correlation of 0.789 (Spearman's rho) with DHH participants' judgments in our PREFERENCE-2017 study data.

—**Weighted Word Error Rate (WWER)** [43]: This metric considers each word to have importance weights (much like our word importance sub-score) such that when a word is recognized an appropriate penalty defined by the importance of the word is assigned to the error. We observed a correlation of 0.742 (Spearman rho) with DHH participants' judgements in our PREFERENCE-2017 study data.

—**Weighted Keyword Error Rate (WKER) and Keyword Error Rate (KER)** [43]: Similar to the WWER metric, this metric penalizes errors based on the importance of the word in the text. However, instead of weighting all the words based on the importance score, this metric follows the strategy of weighting only the keywords while all of the other non-keywords are simply ignored during evaluation. In contrast, KER weighs all the keywords with the score of 1 ignores all the non-keywords. On evaluating the performance of these metrics, the KER metric performed slightly higher (rho = 0.757) than the WKER metric (rho = 0.727), in its correlation with the DHH participants' judgements in our PREFERENCE-2017 study data.

In summary, our ACE metric had a correlation of rho = 0.742 with the DHH users' judgements of text quality in the PREFERENCE-2017 study, and our new ACE2 metric had a correlation of rho = 0.866. Thus, ACE2 had the highest correlation of any metric in this comparative analysis. For comparison, the ubiquitous Word Error Rate (WER) metric was found to have a Spearman's rho correlation of 0.108 with the DHH participants' judgements in this same dataset.

8 PHASE 4: USER-BASED EVALUATION OF ACE AND ACE2

As our final summative evaluation, we now look to compare the performance of our original ACE metric with the newly proposed ACE2 metric, on a new dataset, which we have collected in a new study with DHH participants. Our goal is to definitively compare our old and new versions of this metric, as to their efficacy in predicting the quality of caption text for DHH users.

To set up this study, we followed a similar design as in our original PREFERENCE-2017 study (described in Section 5.5). Thus, we shall refer to this new study as "PREFERENCE-2018" for the remainder of this article. Specifically, we presented DHH participants with automatically generated caption texts based on an audio recording of a two-person conversation, and we asked participants to provide their opinion as to the understandability of each caption.

As in the PREFERENCE-2017 study, participants were able to see the reference text (the "intended" message of what the human was saying, as transcribed accurately by a human), and participants were able to consider this, when they were judging the quality of the automatic captions they were asked to evaluate.

Section 8.1 describes our methodology for designing the stimuli for this new study, Section 8.2 discusses the set-up and conduct of the study, and Section 8.3 provides the results of our evaluation, which compares the ACE and ACE2 metrics.

8.1 Designing Stimuli (PREFERENCE-2018 Study)

The stimuli text that we showed to our DHH participants were generated by transcribing speech recordings of a conversation between two hearing people, using a modern ASR system. The recordings were the snippets from the aforementioned Switchboard corpus, which contains recordings of over 2,400 pairs of strangers having a casual conversation over a telephone line.

In our PREFERENCE-2017 study, we had generated stimuli by "engineering" realistic recognition errors in caption texts, to compare the WER metric and the ACE metric. The idea had been to select a pair of text with the same WER score but with different ACE scores (one high and other low), so that when the two captions were presented to the users side-by-side, we could see which caption participants preferred. Subsequently, we used the data from that original PREFERENCE-2017 study in Phases 2 and 3 of our research.

Although that previous study design had enabled us to gather a useful dataset of DHH participants' judgements, for our new PREFERENCE-2018 study, we wanted to use more naturalistic set of caption texts for evaluation. Thus, rather than "engineering" the quality of a caption text (by selecting a subset of possible ASR errors from a larger set of possible ASR errors, as we had done in the prior study), we wanted to directly utilize the output from ASR systems as stimuli for our PREFERENCE-2018 study. This meant that our prior strategy of fixing the score from one metric and varying the score from the other metric was no longer possible, since we were limited to using the exact output from ASR systems.

Thus, we automatically processed the recordings from the Switchboard corpus using a variety of ASR systems: Google's Cloud Speech, IBM's Watson Speech to Text, and Sphinx ASR.[5] Given the output from these ASR systems, we had to select the specific pairs of caption texts to show in the study. We therefore gave preference to selecting pairs of ASR hypothesis texts from different ASR systems that had greater differences (absolute value) between their ACE and ACE2 score. This approach to stimuli selection ensured that we had the potential to compare the correlations between the human subjective scores and each of our metrics. We also selected a mix of texts with low, medium, and high ACE and ACE2 scores, to introduce sufficient variability for the correlation measure. With this approach, we selected 180 stimuli texts for our study; the average WER of our resulting stimuli in the study was 26%.

8.2 User Study Setup (PREFERENCE-2018 Study)

We conducted an hour-long study where each participant was presented with 60 automatic caption text outputs (30 pairs, displayed side-by-side), along with the reference text displayed above the pair (as shown in Figure 16). Before the start of the study, each participant was shown an *introduction video* where they were introduced to the task: they would see captions with errors in them. The participants were asked to provide a subjective judgment on the understandability of each caption text, by indicating their choice on a ten-level scale (frown face to smiley face) with endpoints labeled as "Useless" and "Useful." Aside from sharing the same reference text, the two ASR output texts displayed side-by-side did not share any specific metric value. (This was a notable difference from our previous PREFERENCE-2017 study design, where the two texts displayed side-by-side had identical WER scores but contrasting ACE scores.)

To recruit participants for the study, we reached out to students from the Rochester Institute of Technology and other people who were DHH in surrounding the Rochester community. We collected data from 12 DHH participants (age distribution with mean = 21.67 and standard deviation= 2.53), which included 7 men and 5 women. Among our participants, 6 people identified themselves as Deaf, 3 people as deaf, and remaining as hard-of-hearing. To evaluate the literacy skill of participants, we used Wide Range Achievement Test 4th Edition (WRAT4[6]). In the study, the participants reported an average WRAT score of 70.83 (standard deviation = 11.33); this is comparable to average scores received by DHH users in this study in the past [7], which has been generally reported

[5]Cloud Speech: https://cloud.google.com/speech/; Watson Speech to Text: https://www.ibm.com/watson/services/speech-to-text/; and CMUSphinx: https://cmusphinx.github.io (with standard acoustic and language models).
[6]http://www.pearsonclinical.com/education/products/100001722/wide-range-achievement-test-4–wrat4.html.

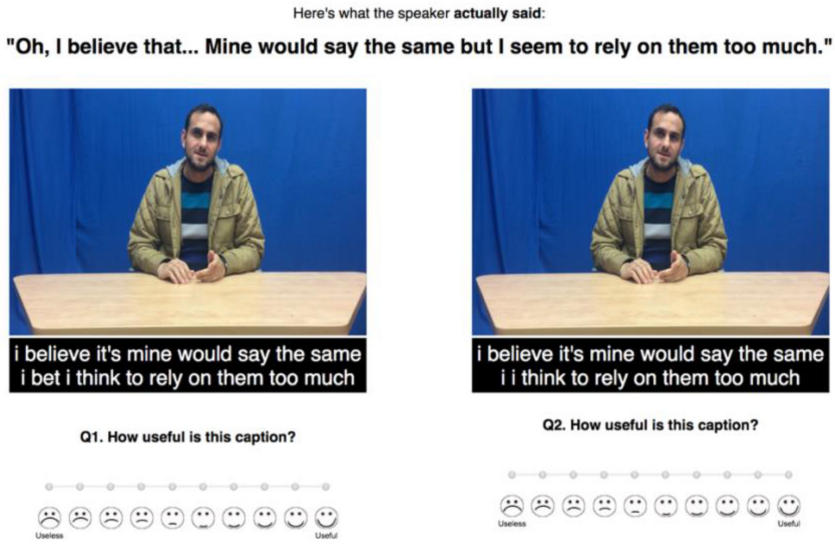


Fig. 16. Screenshot from the study to measure understandability of caption-text automatically generated by ASR.

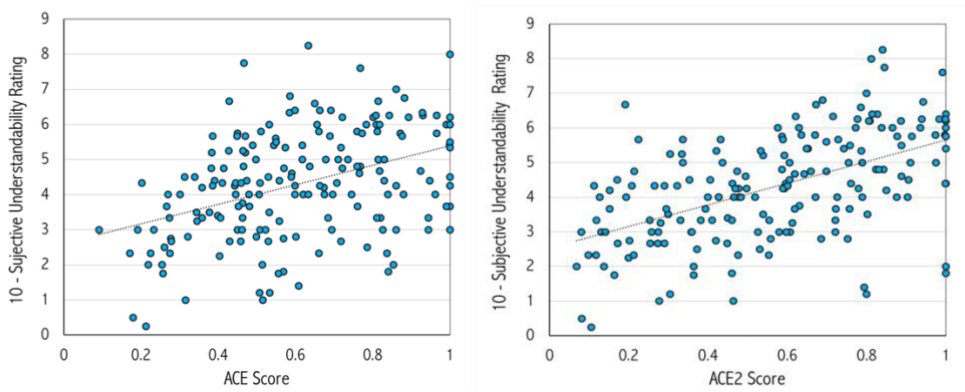


Fig. 17. Analysis of the ACE and ACE2 metric score with the participants' understandability rating.

to be below the national average score of 100 [47]. Last, all of the participants reported that they were familiar with the use of captioning technology, and they regularly used captioning when watching television programming.

8.3 Results and Discussion

In the study, we collected a total of 720 responses from our 12 participants. We created three versions of the stimuli set, with each set containing 60 different caption-text outputs. Each participant in the study judged one version of the stimuli set.

Figure 17 visualizes a linear correlation best-fit line for the relationship between the: (a) participants' subjective judgment rating and ACE scores of each caption-text stimulus and (b) participants' subjective judgment rating and ACE2 scores of each caption-text.

As discussed in Section 4.1, the goal of this study was to investigate our Hypothesis H3, which considered whether the preference scores collected from DHH participants correlated more strongly with the ACE2 metric, as compared to how strongly their correlated with the ACE metric. Therefore, we began our analysis by computing the Spearman Rho values to measure: (a) the correlation between the DHH participants' judgements and the ACE score (r_{ace}) and (b) the correlation between the DHH participants' judgements and the ACE2 score (r_{ace2}). The correlation coefficients were found to be: r_{ace2} = (0.5519, p-value < 0.0001) and r_{ace} = (0.3927, p-value < 0.0001). Figure 17 shows the corresponding correlation graph.

To investigate Hypothesis H3, we next performed significance difference testing on the correlations between r_{ace} and r_{ace2}. Specifically, we performed Fisher r-to-z transformation to calculate the difference. We found a significant difference between the two coefficients (z-score = 1.818, 1-tail p-value < 0.05). Thus, Hypothesis H3 was supported: If we present ASR output texts with varying ACE and ACE2 scores to DHH participants, then there will be a significantly higher correlation between DHH participants' judgments of text-understandability and the ACE2 metric, as compared to the correlation between DHH participants' judgments of text-understandability and ACE.

Thus, we have found that ACE2 out-performs our original ACE metric for the task of predicting DHH users' subjective judgement of the quality of a caption text. However, as we will discuss in Section 8, we observed a drop in the correlation scores (r_{ace} and r_{ace2}) as compared to the earlier Phase 1 study.

9 CONCLUSIONS AND FUTURE WORK

The long-term goal of our research is to investigate the use of ASR technology to provide captioning services for DHH users, especially in real-time contexts such as meetings with hearing colleagues. Current metrics used for evaluating (and sometimes optimizing) the performance of the ASR systems rely on counting the number of recognition errors without regard to what the errors are and where they occur in the transcription. However, this approach of evaluating the performance of the ASR systems had previously been shown to be loosely connected with the actual human-subject opinion on various application settings.

In this work, we wanted to investigate the design and evaluation of new metrics for predicting the quality of an ASR system, to determine whether its output would produce understandable captions for DHH users. Thus, one contribution of this work is in predicting the effect of recognition errors in understanding a caption-text. More specifically, there are two main factors that we found useful in predicting the impact of an error in caption understandability:

- Word Importance Measure: Importance of the error word in the caption text.
- Semantic Distance Measure: Semantic deviation between the error word displayed in the caption text and the actual word that had been spoken.

As a part of our analysis, we have examined various approaches to estimate these measures for predicting the impact of errors in a caption-text. Further, measuring the overall quality of a text based on individual error impact scores required additional analysis on the efficacy of various methods for combining these individual error impact scores, to produce a single aggregate score of the quality of an entire sentence.

Our Phase 1 research had proposed a (baseline) automatic caption evaluation metric, called the ACE metric, and compared the performance of our ACE metric with the WER metric in a caption understandability prediction task, when used in a business meeting scenario. In our side-by-side comparison of pairs of ASR generated caption texts with identical WER, we found: (a) subjective preference of the users on the caption quality judgements from the ACE and, (b) significantly

higher correlation of the DHH users' subjective understandability rating with ACE score as compared to the WER scores.

In our Phase 2 research, we addressed some limitations of the original ACE metric and empirically evaluated various possible improvement strategies. These improvement strategies were evaluated based on the subjective understandability scores collected from the DHH users in PREFERENCE-2017 study conducted in our Phase 1 research. This led to the development of the new metric, called the ACE2 metric, and in Phase 3 of our research, we compared ACE and ACE2 to previously published metrics from other researchers for evaluating ASR text quality.

Finally, in Phase 4 of our research, we collected additional subjective preference data from DHH participants in a new study. However, to diversify the evaluations, the captions used in this study were generated from a more informal two-person conversational speech (compared to the previous fake business meeting setup), and the output of various commercial ASR systems was used to produce stimuli directly, without engineering stimuli to contain specific subsets of ASR errors, as in our prior 2017 study.

Ultimately, our findings reveal that the users' subjective evaluation of the quality of captions is correlated with the number of errors in text, and we have described and evaluated a metric that can be used by future researchers for evaluating the suitability of ASR systems for generating captions for DHH users. Such a metric can be used as an initial investigation of caption quality under various environmental conditions or speakers, and it could be used to compare various ASR systems for this application—prior to conducting a study with DHH users. We also see potential for such metrics to be used to drive the development of ASR-based captioning systems, rather than the use of currently popular metrics, such as WER, which had very little correlation to DHH users' judgements of text quality.

9.1 Limitations of this Research and Plans for Future Work

While our new ACE2 metric has outperformed pre-existing metrics at predicting the subjective judgements of DHH users as to the quality of caption texts, we still see room for improvement: The correlation coefficients between ACE2 and DHH users' judgements in our final summative PREFERENCE-2018 study were modest. Besides the differences in the experimental setup of the two studies, our PREFERENCE-2018 study has other notable differences compared to the PREFERENCE-2017 study, which could further explain the drop in the correlation numbers. For instance, the stimuli from the PREFERENCE-2017 study includes full-length sentences that are part of a business meeting context. In comparison, the stimuli from the PREFERENCE-2018 study are excerpts from a two-person telephone conversation that aren't necessarily grammatical nor topical, demanding greater cognitive load during comprehension. In the future, we would like to conduct a more controlled study to test this hypothesis more accurately.

Furthermore, the ACE framework currently considers one error at a time (fixing all other errors) to measure each error's individual impact in the understandability of the text. While for some applications, it may be useful to identify the individual contribution of specific errors in a text, in other contexts, it may be beneficial to consider models of error impact that represent more complex interactions among multiple errors in a text. To support such research, it would be valuable to collect more qualitative data on the impact of errors in text comprehension studies with DHH participants. Such a large data resource could enable researchers to learn more complex inferences on impact due to multiple errors, without resorting to a more "controlled" single-error-based analysis.

A larger study with a more diverse WER scores and participant age-groups is also desirable for a stronger, more representative analysis. Although our PREFERENCE-2018 study was a follow-up study to the PREFERENCE-2017 study, it could have benefitted from a larger sample of users. Moreover, as discussed in Section 5.4, this study focused on caption texts at a range of WER that

arise in contexts that are challenging for ASR systems, but future work could examine caption text at lower WER. Moreover, all of our metrics (and their sub-components like word importance models and semantic distance models) are designed and evaluated on English texts. Additional studies need to be done to understand their effectiveness for other languages.

It should also be noted that the performance of each of the metrics have been evaluated based on their ability to predict the quality of the transcription of a full utterance unit. While this assumption is reasonable in a conversational setting, where the conversation is more dyadic, this approach would be less applicable if captioning technology were to be used to support a single-speaker channel, such as in classroom or live lecture. Thus, in future work, we plan do a formal evaluation of the metric performance in measuring the quality of longer texts spans rather than individual conversational utterance units. Essentially, it would be necessary to perform automatic segmentation of the longer text transcript generated by the ASR and use the individual sentences/utterances identified in this longer text as the basis for evaluation.

Further, we foresee additional avenues for boosting the performance of the ACE2 metric, through additional research on models of the importance of word in a text and of the semantic distance between error words and the intended word. More specifically, in future work, we plan to investigate other supervised approaches for word-importance prediction, e.g. by training statistical models based on word-importance information collected directly from DHH users. We will also investigate other semantic-distance models, e.g., based on additional semantic features (such as POS-tags, sentiment, polarity of words, etc.) or by identifying vector representations of words that are better suited to calculating semantic distance, e.g., Reference [44].

This work will be conducted as part of our broader research goal of improving ASR-based captioning for DHH users, as described in Section 1.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

ACKNOWLEDGMENTS

We are grateful to Peter Yeung, Abraham Glasser, Larwan Berke, and Christopher Caulfield, who assisted with the data collection for this study. We also thank our collaborators Michael Stinson, Lisa Elliot, Donna Easton, and James Mallory. Larwan Berke and Christopher Caulfield, our colleagues at RIT, provided us the stimuli videos described in Section 5.4.

REFERENCES

[1] Tom Apone, Marcia Brooks, and Trisha O'Connell. 2010. Caption accuracy metrics project. Caption viewer survey: Error ranking of real-time captions in live television news programs. WGBH National Center for Accessible Media Boston. Retrieved from https://dcmp.org/learn/static-assets/nadh300.pdf.

[2] Keith Bain, Sara H. Basson, and Mike Wald. 2002. Speech recognition in university classrooms: Liberated learning project. In *Proceedings of the ACM Conference on Assistive Technologies (ASSETS'02)*. 192–196. https://doi.org/10.1145/638249.638284

[3] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgements. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. 65–72.

[4] Jon P. Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. 2017. The CHiME challenges: Robust speech recognition in everyday environments. In *New Era for Robust Speech Recognition, Exploiting Deep Learning*. Springer, 327–344. https://doi.org/10.1007/978-3-319-64680-0_14

[5] Nathalie N. Bélanger and Keith Rayner. 2013. Frequency and predictability effects in eye fixations for skilled and less-skilled deaf readers. *Visual Cogn.* 21, 4 (2013), 477–497.

[6] Larwan Berke, Christopher Caulfield, and Matt Huenerfauth. 2017. Deaf and hard-of-hearing perspectives on imperfect automatic speech recognition for captioning one-on-one meetings. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'17)*. 155–164. https://doi.org/10.1145/3132525.3132541

[7] Larwan Berke, Sushant Kafle, and Matt Huenerfauth. 2018. Methods for evaluation of imperfect captioning tools by deaf or hard-of-hearing users at different reading literacy levels. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. 91:1–91:12. http://doi.acm.org/10.1145/3173574.3173665

[8] Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'07)*.

[9] Matjaž Debevc, Danijela Milošević, and Ines Kožuh. 2015. A comparison of comprehension processes in sign language interpreter videos with or without captions. *PLoS ONE* 10, 5 (2015), e0127577.

[10] Ana-Belén Domínguez and Jesus Alegria. 2009. Reading mechanisms in orally educated deaf adults. *J. Deaf Studies Deaf Educat.* 15, 2 (2009), 136–148.

[11] Ana-Belén Domínguez, María-Soledad Carrillo, Maria del Mar Perez, and Jesus Alegría. 2014. Analysis of reading strategies in deaf adults as a function of their language and meta-phonological skills. *Res. Dev. Disabil.* 35, 7 (2014), 1439–1456.

[12] J. R. Duffy and T. G. Giolas. 1971. *The Effect of Word Predictability on Sentence Intelligibility*. Technical report, Submarine Medical Research Laboratory. doi>10.21236/AD0746118.

[13] Matthew W. G. Dye, Peter C. Hauser, and Daphne Bavelier. 2009. Is visual selective attention in deaf individuals enhanced or deficient? The case of the useful field of view. *PLoS ONE* 4, 5 (2009), e5640.

[14] Lisa Elliot, Michael Stinson, James Mallory, Donna Easton, and Matt Huenerfauth. 2016. Deaf and hard of hearing individuals' perceptions of communication with hearing colleagues in small groups. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'16)*. ACM, New York, NY, 271–272. https://doi.org/10.1145/2982142.2982198

[15] Benoît Favre, Kyla Cheung, Siavash Kazemian, Adam Lee, Yang Liu, Cosmin Munteanu, Ani Nenkova, Dennis Ochei, Gerald Penn, Stephen Tratz, Clare R. Voss, and Frauke Zeller. 2013. Automatic human utility evaluation of ASR systems: Does WER really predict performance? In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH'13)*. 3463–3467. Retrieved from http://www.isca-speech.org/archive/interspeech_2013/i13_3463.html.

[16] Maria Federico and Marco Furini. 2012. Enhancing learning accessibility through fully automatic captioning. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility (W4A'12)*. 40. https://doi.org/10.1145/2207016.2207053

[17] Ira R. Forman, Ben Fletcher, John Hartley, Bill Rippon, and Allen Wilson. 2012. Blue herd: automated captioning for videoconferences. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'12)*. 227–228. https://doi.org/10.1145/2384916.2384966

[18] Stefan L. Frank and Roel M. Willems. 2017. Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Lang. Cogn. Neurosci.* 32, 9 (2017), 1192–1203.

[19] Nestor Garay-Vitoria and Julio Abascal. 2006. Text prediction systems: a survey. *Univ. Access Info. Soc.* 4, 3 (2006), 188–203.

[20] John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees. 2000. The TREC Spoken Document Retrieval Track: A Success Story. In *Proceedings of the 6th International Conference Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) (RIAO'00)*. 1–20.

[21] Yashesh Gaur, Walter S. Lasecki, Florian Metze, and Jeffrey P. Bigham. 2016. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th Web for All Conference (W4A'16)*. 23:1–23:8. https://doi.org/10.1145/2899475.2899478

[22] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'92)*, Vol. 1. IEEE, 517–520.

[23] David Grangier, Alessandro Vinciarelli, and Hervé Bourlard. 2003. *Information Retrieval on Noisy Text*. Technical Report. IDIAP.

[24] Robert M. Gray. 2011. *Entropy and Information Theory*. Springer Science & Business Media.

[25] Sharmistha S. Gray, Daniel Willett, Jianhua Lu, Joel Pinto, Paul Maergner, and Nathan Bodenstab. 2014. Child automatic speech recognition for US English: Child interaction with living-room-electronic-devices. In *Proceedings of the 4th Workshop on Child, Computer and Interaction (WOCCI'14)*. 21–26. Retrieved from http://www.isca-speech.org/archive/wocci_2014/wc14_021.html.

[26] Rebecca Perkins Harrington and Gregg C. Vanderheiden. 2013. Crowd caption correction (CCC). In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'13)*. 45:1–45:2. https://doi.org/10.1145/2513383.2513413

[27] Nobuyasu Itoh, Gakuto Kurata, Ryuki Tachibana, and Masafumi Nishimura. 2014. A metric for evaluating speech recognition accuracy based on human perception model. In *Proceeding of INTERSPEECH*.

[28] Dorothy W. Jackson, Peter V. Paul, and Jonathan C. Smith. 1997. Prior knowledge and reading comprehension ability of deaf adolescents. *J. Deaf Studies Deaf Educat.* (1997), 172–184.

[29] Sushant Kafle, Matt Huenerfauth. 2016. Effect of speech recognition errors on text understandability for people who are deaf or hard of hearing. In *Proceedings of the 7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT'16)*. 20–25.

[30] Sushant Kafle and Matt Huenerfauth. 2017. Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'17)*. 165–174. https://doi.org/10.1145/3132525.3132542

[31] Saba Kawas, George Karalis, Tzu Wen, and Richard E. Ladner. 2016. Improving real-time captioning experiences for deaf and hard of hearing students. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'16)*. 15–23. https://doi.org/10.1145/2982142.2982164

[32] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Retrieved from http://arxiv.org/abs/1412.6980.

[33] Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *Eur. J. Cogn. Psychol.* 16, 1–2 (2004), 262–284.

[34] Raja S. Kushalnagar, Walter S. Lasecki, and Jeffrey P. Bigham. 2014. Accessibility evaluation of classroom captions. *Trans. Access. Comput.* 5, 3 (2014), 7:1–7:24. https://doi.org/10.1145/2543578

[35] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST'12)*. ACM, New York, NY, 23–34. https://doi.org/10.1145/2380116.2380122

[36] Walter S. Lasecki and Jeffrey P. Bigham. 2014. Real-time captioning with the crowd. *Interactions* 21, 3 (2014), 50–55. https://doi.org/10.1145/2594459

[37] Xin Lei, Andrew W. Senior, Alexander Gruenstein, and Jeffrey Sorensen. 2013. Accurate and compact large vocabulary speech recognition on mobile devices. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH'13)*. 662–665. http://www.isca-speech.org/archive/interspeech_2013/i13_0662.html.

[38] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach. 2014. An overview of noise-robust automatic speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 22, 4 (2014), 745–777. https://doi.org/10.1109/TASLP.2014.2304637

[39] John L. Luckner and C. Michele Handley. 2008. A summary of the reading comprehension research undertaken with students who are deaf or hard of hearing. *Amer. Ann. Deaf* 153, 1 (2008), 6–36.

[40] Iain A. McCowan, Darren Moore, John Dines, Daniel Gatica-Perez, Mike Flynn, Pierre Wellner, and Hervé Bourlard. 2004. *On the use of Information Retrieval Measures for Speech Recognition Evaluation*. Technical report, IDIAP.

[41] Taniya Mishra, Andrej Ljolje, and Mazin Gilbert. 2011. Predicting human perceived accuracy of ASR systems. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH'11)*. 1945–1948. Retrieved from http://www.isca-speech.org/archive/interspeech_2011/i11_1945.html.

[42] Andrew Cameron Morris, Viktoria Maier, and Phil D. Green. 2004. From WER and RIL to MER and WIL:Improved evaluation measures for connected speech recognition. In *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP'04)*. Retrieved from http://www.isca-speech.org/archive/interspeech_2004/i04_2765.html.

[43] Hiroaki Nanjo and Tatsuya Kawahara. 2005. A new ASR evaluation measure and minimum bayes-risk decoding for open-domain speech understanding. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*. 1053–1056. https://doi.org/10.1109/ICASSP.2005.1415298

[44] Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. arXiv preprint arXiv:1605.07766.

[45] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computation Linguistics (ACL'02)*. ACL, 311–318.

[46] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'15)*. IEEE, 5206–5210.

[47] LeAdelle Phelps, Barbara Branyan. 1990. Academic achievement and nonverbal intelligence in public school hearing-impaired children. *Psychol. Schs.* 27 (1990), 210–217.

[48] Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* 124, 3 (1998), 372.

[49] Keith Rayner, Xingshan Li, Barbara J Juhasz, and Guoli Yan. 2005. The effect of word predictability on the eye movements of Chinese readers. *Psychonom. Bull. Rev.* 12, 6 (2005), 1089–1093.

[50] Keith Rayner, Erik D. Reichle, Michael J. Stroud, Carrick C. Williams, and Alexander Pollatsek. 2006. The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychol. Aging* 21, 3 (2006), 448.

[51] Keith Rayner, Timothy J. Slattery, Denis Drieghe, and Simon P. Liversedge. 2011. Eye movements and word skipping during reading: Effects of word length and predictability. *J. Exper. Psychol.: Hum. Percept. Perform.* 37, 2 (2011), 514.

[52] Marek Rei. 2017. Semi-supervised multitask learning for sequence labeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, Vol. 1. 2121–2130. https://doi.org/10.18653/v1/P17-1194

[53] Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2012. TED-LIUM: An automatic speech recognition dedicated corpus. In *Proceedings of the Intern LREC*. 125–129.

[54] Asad Sayeed, Stefan Fischer, and Vera Demberg. 2015. Vector-space calculation of semantic surprisal for predicting word pronunciation duration. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Vol. 1. 763–773.

[55] Michael S. Stinson, Pamela Francis, Lisa B. Elliot, and Donna Easton. 2014. Real-time caption challenge: C-print. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers and Accessibility (ASSETS'14)*. 317–318. https://doi.org/10.1145/2661334.2661337

[56] Agnieszka Szarkowska, Izabela Krejtz, Zuzanna Klyszejko, and Anna Wieczorek. 2011. Verbatim, standard, or edited? Reading patterns of different captioning styles among deaf, hard of hearing, and hearing viewers. *Amer. Ann. Deaf* 156, 4 (2011), 363–378.

[57] Hironobu Takagi, Takashi Itoh, and Kaoru Shinkawa. 2015. Evaluation of real-time captioning by machine recognition with human support. In *Proceedings of the 12th Web for All Conference (W4A'15)*. 5:1–5:4. https://doi.org/10.1145/2745555.2746648

[58] Yuan Tang. 2016. TF.Learn: TensorFlow's High-level Module for Distributed Machine Learning. Retrieved from http://arxiv.org/abs/1612.04251.

[59] David R. Traum and Peter A. Heeman. 1996. Utterance units in spoken dialogue. In *Proceedings of the Workshop on Dialogue Processing in Spoken Language Systems*. Springer, 125–140.

[60] Mike Wald. 2011. Crowdsourcing correction of speech recognition captioning errors. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility (W4A'11)*. 22. https://doi.org/10.1145/1969289.1969318

[61] Ye-Yi Wang, Alex Acero, and Ciprian Chelba. 2003. Is word error rate a good indicator for spoken language understanding accuracy. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03)*. IEEE, 577–582.

[62] J. P Woodard and J. T. Nelson. 1982. An information theoretic measure of speech recognition performance. In *Proceedings of the Workshop on Standardization for Speech I/O technology*.

[63] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2016. *Achieving Human Parity in Conversational Speech Recognition*. Retrieved from http://arxiv.org/abs/1610.05256.

[64] Joong-O. Yoon and Minjeong Kim. 2011. The effects of captions on deaf student's content comprehension, cognitive load, and motivation in online learning. *Amer. Ann. Deaf* 156, 3 (2011), 283–289.

Received April 2018; revised April 2019; accepted April 2019